Reducing preference reversals: the role of preference imprecision and non-transparent methods

Pinto-Prades, José Luis; Sánchez-Martínez, Fernando Ignacio ; Abellán-Perpiñán, José María; Martínez-Pérez , Jorge E.

Published in:
Health Economics

DOI:
10.1002/hec.3772

Publication date:
2018

Document Version
Author accepted manuscript

REDUCING PREFERENCE REVERSALS: THE ROLE OF PREFERENCE IMPRECISION AND NON-TRANSPARENT METHODS

Running head: Preference reversals, imprecision and non-transparent methods

Keywords:

Preference reversals, preference imprecision, matching, choice, health measurement, Standard Gamble.

Words: 7939

Tables: 7

Figures 3

Authors:

José Luis Pinto Prades – Universidad de Navarra, Spain and Glasgow Caledonian University, UK.

Fernando Ignacio Sánchez Martínez* – Universidad de Murcia, Spain

José María Abellán Perpiñán – Universidad de Murcia, Spain

Jorge E. Martínez Pérez – Universidad de Murcia, Spain

*** Corresponding author**

Fernando Ignacio Sánchez Martínez

Facultad de Economía y Empresa. Campus de Espinardo. Universidad de Murcia.

30100 – Murcia (Spain)

Phone: +34 868 88 37 40

Fax: +34 868 88 37 45

e-mail: fernando@um.es

This research was funded by the Spanish Ministerio de Economía y Competitividad (Projects ECO2013-43526-R / ECO2013-48631-P).

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

ABSTRACT

Preferences elicited with matching and choice usually diverge (as characterized by preference reversals), violating a basic rationality requirement, namely, procedure invariance. We report the results of an experiment that shows that preference reversals between matching (Standard Gamble in our case) and choice are reduced when the matching task is conducted using non-transparent methods. Our results suggest that techniques based on non-transparent methods are less influenced by biases (i.e. compatibility effects) than transparent methods. We also observe that imprecision of preferences influences the degree of preference reversals. The preference reversal phenomenon is less strong in subjects with more precise preferences.

1. INTRODUCTION

Resource allocation decisions in health are often justified on the basis that they reflect societal preferences. In practice, those preferences are reflected in the valuation of health states obtained from techniques like the Standard Gamble or the Time Trade-Off. Values produced by these techniques ought to comply with some rationality assumptions. One of them is procedure invariance: normatively equivalent procedures for assessing preferences should give rise to the same preference order. Unfortunately, there is evidence that this principle is violated in the elicitation of preferences for health states. One example of violation of procedure invariance is the well-known preference reversal phenomenon. Preferences change depending on the method (matching or choice[1]) used to elicit them. The main objective of this paper is to understand better the mechanisms that generate preference reversals and how to avoid/reduce that phenomenon. More specifically, we study the role of different preference elicitation methods and the role of preference imprecision in the explanation of the preference reversals.

Violations of procedure invariance have been observed in the health domain in the two main methods used in health economics, namely, Time Trade-Off (Sumner & Nease, 2001) and Standard Gamble (Oliver, 2013a; 2013b). There is also evidence suggesting that utilities elicited with choice experiments and utilities elicited with matching methods (Time Trade-Off and Standard Gamble) are different (Bansback, Brazier, Tsuchiya, & Anis, 2012; Robinson, Spencer, Pinto-Prades, & Covey, 2016). Choice and matching seem to produce different results. According to Oliver (2013a, abstract)

[1] In choice the decision maker selects an option from an offered set of two or more alternatives. In matching the decision maker is required to set the value of some variable in order to achieve equivalence between options.

"those violations pose a challenge to health economics, where choice and valuation methodologies often are used interchangeably". The literature suggests (Fischer, Carmon, Ariely, & Zauberman, 1999; Bostic, Herrstein, & Luce, 1990; Loomes & Pogrebna, 2016) that matching methods based on non-transparent sequences of choices can produce results more consistent with choices. However, there is very little evidence about the role that these methods can play in avoiding preference reversals in the domain of health outcomes. Attema and Brouwer (2013) test the internal consistency of choice and matching tasks, using Time Trade-Off, and conclude that choice tasks – more precisely, choice-based matching (CBM) tasks, lead to more consistent results than direct matching tasks. One objective of this study is to determine if there are elicitation methods that reduce or eliminate the discrepancy between matching and choice in the domain of the evaluation of health outcomes. It has also been suggested (MacCrimmon & Smith, 1986; Butler & Loomes, 2007; Loomes & Pogrebna, 2016) that imprecision in preferences may also explain the degree of preference reversals. For this reason, another objective of this study is to test the role of preference imprecision in the preference reversal phenomenon.

We present the results of an experiment that compares choices and several matching procedures. All these matching methods are choice-based since they match two options using sequences of choices. Our main finding is that CBM methods which hide the final goal of the sequence from respondents (what we call "non-transparent" methods) reduce the rate of preference reversals. That is, when subjects do not see that each choice is part of a sequence aimed at establishing indifference between options, the number of preference reversals is reduced. We suggest that methods like the Time Trade-Off or the Standard Gamble can be improved using non-transparent sequences of

choices. We also find that preference reversals are related to preference imprecision. In the next section, we review the explanations provided in the literature about the choice-matching discrepancy. Sections 3 and 4 contain the methodology and the results, respectively, of our study. The paper ends with the conclusions.

2. THE CHOICE-MATCHING DISCREPANCY

2.1. The phenomenon

The choice-matching discrepancy was first observed using monetary gambles (Lichtenstein & Slovic, 1971). In stylized form, the phenomenon can be explained as follows. Subjects are presented with two lotteries, L1 and L2. Assume, that L1=(Q,m;0) and L2=(q,M;0), where m and M are the best outcomes of L1 and L2 respectively while q and Q are the corresponding probabilities. M is significantly larger than m and q is significantly lower than Q. The expected value of L1 is not too different from L2 (maybe a little bit larger for L2 since it is riskier than L1). L1 is called the P-bet since it offers a high chance of winning some positive outcome and L2 is called the $-bet since it offers the possibility of winning a larger prize. Subjects are asked to state their preference between the two lotteries with two different tasks: monetary equivalence (a matching task) and direct choice. In the monetary equivalence task, subjects have to state a certain monetary amount that has the same value as the lottery. In the choice task, subjects must choose between the P-bet and the $-bet. A reversal may occur when an individual gives a higher monetary equivalent to the $-bet but prefers the P-bet in direct choice or, conversely, when she assigns a higher monetary equivalent to the P-bet but prefers the $-bet in direct choice. Lichtenstein & Slovic's finding, confirmed many times

since then (Seidl, 2002) was that the first type of discrepancy greatly outnumbered the other one.

Slovic, Griffin, and Tversky (1990) studied the potential discrepancy between choice and matching using what they called probability matching. They asked people to state the probability q^* such that subjects were indifferent between the lottery (Q,m;0) and the lottery (q^*,M;0). Interestingly, they did not find preference reversals in this case. Butler and Loomes (2007) found a discrepancy between choice and matching in another type of matching task, that they called Probability Equivalent. In their case, matching was conducted using a "Reference lottery" (p,R;0) (R-bet). This lottery was characterised by having a 0 outcome if unsuccessful, and a payoff larger than the highest payoff in the $-bet if successful ($R$>$M$). Subjects had to state the probabilities in the R-bet such that they were indifferent between the R-bet and the lottery under evaluation (P-bet or $-bet). We will call this method Reference Probability Equivalent (RPE). Butler and Loomes (2007) found the opposite asymmetry to Lichtenstein and Slovic (1971) for monetary equivalents, that is to say, P-bet more favoured in matching than in straight choice.

2.2. Explanation of the phenomenon

Preference reversals have been explained in several ways. One explanation is psychological, since it is based on a principle called *compatibility between stimulus and response*, originally observed by Fitts and Seeger (1953) in sensory tasks. Another explanation is that some people have truly intransitive preferences because, for example, their preferences can be characterized by Regret Theory (Loomes & Sugden, 1982). A different explanation is built on the idea of preference imprecision (MacCrimmon & Smith, 1986; Butler & Loomes, 2007). We analyse in this paper the explanations based on compatibility and imprecision.

2.2.1. Compatibility.

According to this principle, the respondent weights more heavily those characteristics of the stimulus that are more compatible with the response. Several effects have been explained using this general principle. One is *scale compatibility*, according to which, it is the similarity between stimulus and response scales that leads subjects to overweight the compatible attribute. In the study of Lichtenstein and Slovic (1971), the response scale was money in the matching task, so the monetary outcomes of the lotteries were overweighted. This favours the $-bet lottery in matching, since it has the highest monetary price. The finding, in Butler and Loomes (2007), that the P-bet was preferred to the $-bet using RPE could also be explained by *scale compatibility*[2], since probability is the attribute that is used to reach indifference.

A second form of compatibility that may contribute to the choice-matching discrepancy is *strategy compatibility*, which assumes that the strategies that subjects follow in each task are different. In choices, subjects may follow qualitative strategies; for example, a choice may be decided using lexicographic principles or aspiration levels. The implication is that, in choice, subjects focus mainly on the most important (or prominent) attribute. This is called the *prominence hypothesis* (Tversky, Sattath & Slovic, 1988), which implies that the most important attribute receives a higher weight in choices than in matching. In summary, scale compatibility leads to an overvaluation of the $-bet when matching is conducted using monetary equivalents while choice leads to an overvaluation of the P-bet due to prominence, assuming that probability is the

[2] Butler and Loomes (2007) explain their results in a different way, namely, imprecision of preferences. See 2.2.2.

most important attribute. This leads to the well-known result that P-bet≻$-bet in choice and $-bet≻P-bet in matching.

When matching is conducted using the prominent attribute, the psychological interpretation depends of the relative strength of the two effects. Slovic et al. (1990) suggest that the lack of discrepancy between probability matching and choice in their study can be explained by two effects (scale compatibility and prominence) going in the same direction with the same strength. In matching the P-bet is favoured over the $-bet because of scale compatibility and in choice the P-bet is favoured over the $-bet because probability is the most prominent attribute. Fischer and Hawkings (1993), using choices under certainty (e.g. apartments characterized by price and distance to university campus) also compared choice against matching in the prominent attribute (price) and found that the cheaper apartment was more preferred in choice (55%), than in price matching (31%). They interpret their result as evidence that the effect of prominence was larger than the effect of scale compatibility.

Fischer et al. (1999) provide another explanation for the choice-matching discrepancy: the *Task-goal hypothesis*. According to this, the prominent attribute is weighted more heavily in tasks whose perceived goal is to differentiate between alternatives than in tasks whose goal is to equate alternatives. The reason is that to differentiate only requires to rank-order the alternatives, which is naturally compatible with choosing the alternative that is superior in the prominent attribute. To equate requires making trade-offs between attributes, which is naturally compatible with giving some weight to all the attributes. The implication is that the prominent attribute will receive more weight in response tasks whose perceived goal is to differentiate between alternatives (choice) than in tasks whose perceived goal is to equate between alternatives (matching).

2.2.2. Imprecision

The second explanation of preference reversals is based on the idea that preferences are imprecise for, at least, the type of experimental tasks presented to subjects. This explanation was suggested originally by MacCrimmon and Smith (1986) and it has been explored in Butler and Loomes (2007) and Loomes and Pogrebna (2016). Those papers explain preference reversals using two main assumptions, namely, preferences are imprecise, and "the imprecision interval is positively correlated with the ranges within which dominance is not transgressed" (Butler & Loomes, 2007, p. 290). This implies that in the monetary equivalent task the imprecision interval is wider for the $-bet than for the P-bet but in the RPE task, the imprecision interval is wider for the P-bet than for the $-bet. We will use a numerical example to illustrate the link between preference reversals and imprecision in both cases.

We start with preference reversals between monetary equivalents and choice. Assume that a subject is truly indifferent in choice between the following two lotteries: (0.8, €20; 0) – the P-bet - and (0.2, €90; 0) – the $-bet -. We ask the subject to state the monetary equivalent for each of those two lotteries. If preferences were deterministic and transitive, the monetary equivalent would be exactly the same amount of money for both lotteries. The subject would always provide the same response. However, subjects with imprecise preferences are not sure about the exact monetary equivalent. Sometimes they may give an answer and sometimes a different one. Those answers will be bounded by transparent dominance. The upper bound is higher in the $-bet (€90 in our example) than in the P-bet (€20) while the lower bound is the same (€0). The range within which dominance is not transgressed will be [0, €20] for the P-bet and [0, €90]

for the $-bet. Given the assumption that the imprecision interval is positively correlated with those ranges, the imprecision interval will be wider for the $-bet than for the P-bet. For example, the imprecision interval could be [€10, €16] for the P-bet and [€12, €30] for the $-bet. The imprecision interval will contain larger values for the $-bet than for the P-bet. If this is the case, it will be easier for the subject to state a larger monetary equivalent for the $-bet than for the P-bet even though she is truly indifferent in choice. This will produce the pattern observed, namely, $-bet more favoured in matching than in straight choice.

Let us now see what this model predicts for the choice vs. matching discrepancy using RPE. Using our example, assume that the reference lottery is R:(p, €120; 0). The response (p) will be bound by the range within which dominance is not transgressed. This range is larger for the P-bet [0, 0.8] than for the $-bet [0, 0.2]. Again, under the assumption that the imprecision interval is correlated with those ranges it will be wider for the P-bet than for the $-bet. For example, it could be [0.10, 0.16] for the $-bet and [0.12, 0.30] for the P-bet. If this is the case, it will be easier for the subject to state a larger probability value for the P-bet than for the $-bet in the RPE task even though she is truly indifferent in choice. Applying the same argument as before, matching with RPE would produce the pattern observed by Butler and Loomes (2007), namely P-bet more favoured in matching than in straight choice.

2.3. Classifying choice-based matching methods

2.3.1. Standard matching vs. choice-based matching.

Assume we have two objects (A, B) with two attributes (X, Y), that is, object A is characterised by (X_A, Y_A) and object B by (X_B, Y_B). To estimate a combination of attributes such that A and B are equally attractive, we may fix three of the four attributes and ask

the subject to specify the value of the omitted attribute that makes her indifferent between A and B. For example, the subject has to state the missing value (?) in an open question so that (X_A, Y_A) and (X_B, ?) are equally preferred. We call this standard (or direct) matching. Choice-based matching (CBM) does not reach indifference with an open question but with a sequence of choices. This process generates an interval where the indifference point is located. Assume that (X_A, Y_A) $\succ$ (X_B, Y_B^1) but (X_A, Y_A) $\prec$ (X_B, Y_B^2). We then know that the value of Y_B that will make the two options equally attractive will be in the interval $[Y_B^1, Y_B^2]$. Researchers may take the middle point of the interval as the indifference (matching) point, or they may ask an open question with the matching point constrained by the interval where indifference was located.

2.3.2. Iterative vs. non-iterative choice-based matching

In an iterative CBM method, the choice that the subject is presented depends on her response to a previous choice. For example, assume that the subject says that (X_A, Y_A)$\prec$(X_B, Y_B^1). In an iterative method, the subject would then be offered a choice between (X_B, Y_B^2) and (X_A, Y_A) for $Y_B^2 < Y_B^1$, but she would never be presented with a choice between (X_B, Y_B^2) and (X_A, Y_A) for $Y_B^2 > Y_B^1$. In non-iterative methods, a number of questions are set up in advance, each including different values of Y_B and subjects respond to all of them independently of their responses to previous choices.

In health economics, CBM methods (Time Trade-Off and Standard Gamble) are usually iterative. Non-iterative methods are not very common in health economics, although the *Multiple Price List* method, widely used to elicit risk preferences (Holt & Laury, 2002) is non-iterative. One problem with non-iterative methods is that subjects can give inconsistent responses and an indifference point (or interval) cannot be estimated for those subjects. Finally, depending on the rules to generate the stimuli from one choice

to another we can further split iterative methods into Titration, Bisection, “Ping-pong” and so on. The differences between those methods will be explained later.

2.3.3. Transparent vs. non-transparent choice-based matching.

When subjects can easily observe that there is a link between the choices of a converging sequence, we say that the method is “transparent”. If it is difficult for people to observe this link, we talk about a “non-transparent” method. Let us see an example. Assume we want to estimate the utility of *N* health states. This requires *N* converging sequences of choices in CBM. The usual way of eliciting preferences would be to start with a certain health state (say health state 1), apply the corresponding converging sequence (say #1) until indifference is reached and then move to a different health state and do the same. Since choices of converging sequence #1 keep all attributes constant except the attribute used to match options (e.g. time in Time Trade-Off, probability in the Standard Gamble), some subjects may quickly realise the kind of “game” they are playing. However, an alternative way of eliciting preferences would be to ask subjects to make one choice from the converging sequence #1, then one choice from sequence #2 (a choice corresponding to health state 2), one choice from sequence #3 (a choice corresponding to health state 3), …, one choice from sequence #*N* (a choice corresponding to health state N), before returning to a choice corresponding to sequence #1. In this way, when the subject is presented with the second choice of converging sequence #1, we hope she cannot see that this choice is related to a choice that she made *N* questions previously. This is the method used by Fischer et al. (1999), which they coined as *Hidden Choice-Based Matching* (HCBM). This is a non-transparent method.

While the iterative or non-iterative nature of a CBM method is objective, it is not the same in the case of transparency. For example, in the HCBM method, we can expect the method to be less non-transparent the fewer converging sequences we use. However, we would expect that the majority of people might find it more difficult to observe that the choice belongs to a sequence with HCBM than with the ordinary CBM methods.

2.4. Avoiding the discrepancy

In this paper, we address the issue of how to avoid preference reversals. If we accept the psychological interpretation based on compatibility and prominence, we can predict that some preference elicitation methods will produce more preference reversals than others. If scale compatibility or strategy compatibility were the reasons behind the choice-matching disparity, it would be enough to move from standard matching to CBM to avoid preference reversals. If each of the tasks in CBM is perceived as an independent choice, preferences elicited with CBM and direct choices should converge, since the matching task becomes qualitative and it does not involve the use of a scale to match two options. However, if the Task-Goal is the correct explanation of the discrepancy, it would not be enough to use CBM methods to avoid preference reversals. If CBM are transparent the subject will perceive that the objective of the CBM task is to equate between alternatives while the objective of the choice task is to differentiate between alternatives. To avoid preference reversals subjects should not perceive the objective of the task (equate vs. differentiate). In this paper, we will use different CBM methods that combine the iterative or non-iterative nature of the procedure with its transparency or opacity. If the combination of scale and strategy compatibility explains the choice-matching discrepancy, all methods should eliminate the discrepancy since they are all

choice-based. If the Task-goal is the correct explanation, non-transparent methods will perform better in that regard.

If the explanation based on imprecision in preferences is correct, we have to introduce some mechanism to detect the degree of imprecision in preferences. For this reason, we asked subjects to repeat their choices and valuations in three occasions. According to the imprecision hypothesis we should find more preference reversals in those subjects with more imprecise preferences.

3. METHODS

3.1. Participants

We recruited 250 undergraduate students at the University of Murcia (Spain). Participants were randomly allocated[3] to one of five groups, which differed in the type of elicitation mode used in the CBM procedure. The sessions took place in the Lab of the Faculty of Economics and Business of the University of Murcia under the supervision of members of the research team. Fourteen sessions were held with less than 25 students in each session. Students were paid €15 for their participation. Sessions took about 40 minutes to complete the study.

3.2. Gambles and tasks

Two pairs of lotteries were used (see Table I). In each lottery, one outcome was a chronic health condition described in terms of an EQ-5D-3L[4] health state, and the other outcome

[3] We included all subjects who volunteered for the experiment in a database, and they were allocated to one of the five CBM methods using a random number generator. Since we wanted to have exactly the same number of subjects in each version, the number corresponding to each CBM method was omitted once we had 50 subjects allocated to that procedure. When they introduced their National Identity Card number in the computer, they were allocated to only one of the methods.

[4] The EQ-5D-3L descriptive system includes five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three levels: no problems, some problems and severe problems. State 12231, for instance, describes the condition of an

was immediate death. Lotteries A and C (P-bet) offered the individuals a large probability of being in a bad health state for the rest of their lives, whereas lotteries B and D ($-bet) gave them a low probability of living in a better health state but a higher risk of death. The two sets of paired lotteries (A-B, C-D) had similar expected utilities according to the EQ-5D-3L Spanish tariff (Badia, Roset, Herdman, & Kind, 2001).

INSERT TABLE I

The scenarios described a hypothetical situation where subjects had to choose between two treatments - otherwise, they would die in a few days. Assuming that the reference point was certain death (and assuming utilities of the four health states are positive), each treatment was in the gain domain. Therefore, a P-bet is a treatment that offers a large probability of a small health gain, while the $-bet offers a smaller chance of a bigger health gain. Visual aids were used to represent the probabilities of success and failure in each treatment. An example can be seen in Figure 1 that represents a direct choice between lotteries A and B.

INSERT FIGURE 1

The valuation task consisted of a sequence of choices between each of the lotteries and a reference gamble (R), whose best outcome was full health (more precisely, the best state described by the EQ-5D-3L descriptive instrument) and the worst was death, that is, R: (*p,* 11111; *Death*). We therefore used a Reference Probability Equivalent (RPE) technique, as in Butler and Loomes (2007), where the attribute used to reach the equivalence is the probability in the reference lottery (*p*). An example of a RPE question can be seen in Figure 2.

individual who has no problems in walking nor is anxious or depressed, but who has some problems washing or dressing him/herself and performing usual activities, and who has extreme pain or discomfort too.

INSERT FIGURE 2

The possible values for probability *p* were predetermined before the CBM procedure started. It was necessary to define in advance the values of *p* given that, in non-iterative methods, the subject was asked a predetermined number of questions independently of her responses to previous questions. In order to make matching tasks between the five methods as similar as possible, we decided to adopt the same predetermined values in iterative and non-iterative methods. In each of the four lotteries, nine different values of *p* were used (see Table II). In iterative methods, subjects were only asked a subset of those nine values, while in non-iterative they were asked all the nine values. Indifference was not allowed in straight choices (A vs. B; C vs. D) or in RPE.

INSERT TABLE II

3.3. Choice-based matching methods

Five different types of CBM were used: two were iterative and transparent; one was iterative but non-transparent; one was non-iterative and transparent, and the last one was non-iterative and non-transparent. We describe each of them in turn.[5]

Transparent iterative methods (Bisection and Ping-pong)

The transparent iterative methods we used were Bisection and a modified version of the Ping-pong procedure. In both methods, the first value of the matching parameter was randomly chosen amongst the nine potential values of *p*. Assume it was p_3. This generated two potential intervals where the indifference point had to be located, namely, [0-p_3] and [p_3-p^i_{max}]. For example, in the comparison between lottery (12231,

[5] There was another group where we tried to implement the so-called "PEST" method (Bostic et al., 1990). However, the programming went wrong and the method turned out being almost identical to Ping-pong. The results were very similar to Ping-pong.

0.95; *Death*) and (11111, p_3; *Death*), we had p_3=0.3. If the respondent chose lottery (11111, 0.3; *Death*), we knew that the next value of *p* had to be 0.1 or 0.2. If the respondent chose lottery (12231, 0.95; *Death*), the second stimulus had to be one value from the set {0.4, 0.5, 0.6, 0.7, 0.8, 0.9}.

The difference between both methods (Bisection and Ping-pong) was how they chose the value of *p* in the subsequent question. Assume the subject preferred lottery (12231, 0.95; *Death*) to lottery (11111, 0.3; *Death*). In the Bisection method, the second value of *p* would be the value closest to the middle point of the interval [p_3-p^i_{max}], which, in this case, would be p_7. In the Ping-pong method, the second value of *p* would be located at the other end of the interval opposite p_3. In this example, it would be p_9. In both cases, the process went on until the indifference point was located within one of the ten intervals $[\overline{p^i_L - p^i_U}]$ defined. At this point, the process stopped. The lower limit of the interval (p^i_L) was the highest value of *p* in lottery R (11111, *p*; *Death*) for which the individual preferred the lottery *i* to *R*; and the upper end of the interval (p^i_U) was the lowest value of *p* in lottery R for which the subject preferred the lottery *R* to *i*. For example, if (12231, 0.95; *Death*)$\succ$(11111, 0.6; *Death*) but (12231, 0.95; *Death*)$\prec$(11111, 0.7: *Death*) then p^i_L=0.6 and p^i_U=0.7 so $[\overline{p^i_L - p^i_U}]$ = $[0.6 - 0.7]$.

Hidden Choice-Based Matching method (HCBM)

The non-transparent iterative procedure used was the HCBM proposed by Fischer et al. (1999). The HCBM was applied using the Bisection method but separating the choices regarding each particular lottery by the iterations of other lotteries. Thus, subjects made one choice from the iteration process belonging to lottery A, then one choice for lottery

B, one for lottery C and one for lottery D before returning to the sequence of lottery A.[6] For example, a hypothetical sequence could have been as follows: the first choice between (12231, 0.95; *Death*) and (11111, 0.3; *Death*), the second choice between (11221, 0.3; *Death*) and (11111, 0.09; *Death*), the third choice between (22223, 0.8; *Death*) and (11111, 0.56; *Death*) and the fourth choice between (12221, 0.2; *Death*) and (11111, 0.16; *Death*)[7]. Assuming that, in all cases, the reference lottery was preferred, the fifth to eighth choices were (12231, 0.95; *Death*) vs. (11111, 0.2; *Death*), (11221, 0.3; *Death*) vs. (11111, 0.06; *Death*), (22223, 0.8; *Death*) vs. (11111, 0.32; *Death*) and (12221, 0.2; *Death*) vs. (11111, 0.08; *Death*).

'List' method

This method, like the following one, is non-iterative, that is, subjects had to respond to all possible predetermined choices (p_1^i to p_9^i). The method consisted in a list/table containing all the possible choices for each of the matching sequences, which were displayed in random order.[8] Figure 3 shows a selection of the table that respondents saw.

INSERT FIGURE 3

Random Binary Choice (RBC) method

[6] As Fischer et al. (1999) did in their study, when a sequence converged faster than others, filler choices were added at the end of that sequence to avoid the possibility that there were only one or two sequences that had not converged at the final stages, thus, making the iterative process transparent to the subjects.

[7] In the first sequence of choices, probabilities of the reference lottery (0.3, 0.09, 0.56 and 0.16, respectively) were randomly set amongst the nine potential predetermined values in each case.

[8] Presenting choices in a random order is not the usual way of administering multiple price / choice lists. Our intention was to design a framing which looked almost identical to the following one, RBC, except for the fact that one (List) would be transparent and the other one (RBC) non-transparent.

This method resembled (to some extent) a DCE experiment keeping the "essence" of matching, that is, we obtained the indifference point for each subject and each health state. RBC can be seen as a mixture between List and HCBM. The method is non-iterative (as List), since it presented the nine potential choices to each subject for each of the four lotteries. However, unlike List, the method is non-transparent and it used the same strategy as HCBM to avoid being transparent, namely, choices corresponding to a particular lottery were presented randomly, one at a time and with choices related to other lotteries interspersed between them – as in the HCBM procedure. The difference between HCBM and RBC is that in RBC the subject was asked all nine choices for each lottery while HCBM only asked questions that could not contradict previous choices since it is iterative. As has been pointed out, in non-iterative methods (List and RBC) an indifference interval $[\overline{p_L^i - p_U^i}]$ might not be determined instantly if responses are noisy. A method to deal with this problem will be explained later.

3.4. Structure of the sessions

The sessions began with an introduction to the experiment. The EQ-5D-3L descriptive system was briefly explained to the participants and the four health states involved in the lotteries were shown. Subjects were then asked to rate the four states plus the *Death* state on a visual analogue scale.

After that, subjects had to do two types of tasks: choices between paired lotteries (one P-bet and one $-bet) and valuations of each lottery by means of the RPE technique. In all groups, the first task was to choose between lotteries A and B. Then, subjects in groups 1 (Bisection), 2 (Ping-pong), and 4 (List) were asked to do the RPE task of A and B, followed by the choice between C and D and RPE of C and D. In groups 3 (HCBM) and 5 (RBC), subjects started with the two straight choices (A vs. B and C vs. D) and then

continued with the valuations of the four lotteries in the manner that has been explained. The same scheme was repeated three consecutive times during the session. For groups 1, 2 and 4, the order in which A and B were valued through the RPE method was randomly determined, as was the order between C and D valuations. For groups 3 and 5, the order of appearance of the four lotteries in RPE valuations was set at random.

3.5. Analysis

There are, at least, two ways, of analysing the data of this experiment. One is to analyse the responses of each round separately. We count the number of preference reversals in each round and we see if they vary between methods. It is as if we had three different subsamples, one for each round. However, the preference reversal asymmetry is an individual phenomenon, so it would be good to find a way of characterizing subjects as having or not "true" preference reversal preferences. This is the objective of the second approach. It combines all responses that each subject provided in the three rounds to characterize subjects' preferences. We proceed to explain the two approaches.

3.5.1. Counting Preference Reversals

Within each pair of lotteries *i* and *j*, a choice-matching discrepancy (a preference reversal) exists when $i \succ j$ in a straight choice, but RPE valuations imply that $i \prec j$. Since we did not derive an exact indifference value for the lotteries but only an indifference interval $[\overline{p_L^i - p_U^i}]$, we need to define the procedure to estimate the choice implicit in a RPE task. One option is to assume that the indifference value is the middle point of the indifference interval. Another option is to compare the indifference intervals and assume that the RPE valuation task reveals a preference for *i* or *j* only when the two intervals do not overlap. For instance, when a RPE task implies that $i \succ j$ if $p_L^i > p_U^j$. If p_L^i and p_U^j overlap, no direction of preference can be established.

3.5.2. Classifying subjects.

When preferences are imprecise, some individuals may respond differently when the same question is asked several times at different moments in the course of an experimental session. In our case, subjects may sometimes choose A (C) and sometimes B (D), and the probabilities elicited in RPE very often overlap due to preference imprecision. Taking that into account, we proceed as follows. We use the three straight choices between each pair of lotteries (*i*, *j)* of each subject and we define as "truly preferring one lottery in choice" those who chose one lottery at least twice. We call "strongly consistent" those who chose the same lottery in each of the three rounds and "weakly consistent" those who chose the same lottery twice. This leads to a classification of subjects in four groups depending on whether they were strongly or weakly consistent and whether they prefer lottery *i* or lottery *j*.

It is more complicated to decide when a subject truly prefers one lottery or the other with the valuation (RPE) task. We use the notion of "stochastic indifference" (SI) developed by Loomes and Pogrebna (2016) to do that. SI may be derived from a series of repeated valuation tasks counting the number of times that one option is chosen against the other. For example, assume that the subject has to respond to a series of choices between lottery A:(12231, 0.95; *Death*) and the reference lottery R:(11111, *p*; *Death*), for p = 0.1, 0.2, …, 0.8, 0.9 (see Table II) three times. Assume that the subject has well-defined (deterministic) preferences and she always chooses lottery A for any value of $p \leq 0.6$ and always chooses the R-gamble when $p \geq 0.7$. In that case, the indifference interval would be (0.6, 0.7) and the indifference point could be set in the middle point of this interval, that is, 0.65. This subject would choose lottery A in 18 of the 27 choices between A and R. Assume now that the subject has imprecise (stochastic)

preferences. She always chooses lottery A for $p \leq 0.4$ and always lottery R for p=0.9. For p=0.5 and p=0.6 she chooses lottery A twice and for p=0.7 and p=0.8 she chooses lottery A just once. Overall, she chooses A in 18 of the 27 potential choices. We then say that her stochastic indifference point is also 0.65, as in the deterministic case.

In general, the stochastic indifference point can be estimated as

$$SI(i) = {}^{1}/_{2}\, p_1^i + {}^{1}/_{3}\left(p_n^i - p_{n-1}^i\right)L^i$$

Where p_1^i is the lowest possible value of *p* in the R gamble (i.e. 0.1 for lottery A, 0.03 for lottery B, 0.08 for lottery C and 0.02 for lottery D); $\left(p_n^i - p_{n-1}^i\right)$ is the gap between the values of *p* in each R gamble (which coincides with its lowest value, that is, 0.1, 0.03, 0.08 and 0.02, respectively for lotteries A, B, C and D) and L^i is the number of times that the subject chooses lottery *i* $(0 < L^i < 27)$. In the example just provided, since $L^i = 18$, $\mathrm{SI(A)} = 0.05 + {}^{0.1}/_{3} \times 18 = 0.65$.

The above method can be immediately applied to non-iterative methods, since all subjects had to respond to all (i.e.27) binary choices. In the case of iterative methods subjects did not respond to all 27 binary choices since once the subject responded to one choice, all potential questions involving dominance were excluded. To estimate SI for iterative methods, we assume that when a participant chose the lottery *i* (*j*) against R: (11111, *p*; *Death*) for a certain value of *p* = *p*', she would also have chosen *i* (*j*) for any *p* < *p*'. And, conversely, if a participant chose the reference gamble R, for a given value of *p* = *p*'', she would also have chosen R for any *p*>*p*''. In summary, using the concept of Stochastic Indifference we split the subjects into those who truly preferred one lottery or the other in RPE.

In summary, subjects could belong to one of eight potential groups generated by three two-level factors: a) if they prefer one lottery or the other in choice, b) if they prefer one

lottery or the other in valuation, c) if they are strongly – choose one lottery 3 times - or weakly - choose one lottery 2 times - consistent in choice. The prediction here is that those who are weakly consistent will make more preference reversals than those with better defined preferences.

4. RESULTS

4.1. Round by round analysis

We present the results of each round separately in Tables III to V. The total number of preference reversals was 129 in the first round (Table III), 74 in the second round (Table IV) and 76 in the third round (Table V) so there seems to be an element of learning in our subjects, at least between the first and the second round (the rate of preference reversals falls from 29 % in round 1 to 16 % in rounds 2 and 3). However, in spite of the reduction in the number of preference reversals, they are still highly asymmetric for Bisection, Ping-pong and List in the same direction, namely, the P-bet being more highly preferred in valuation than in choice. This coincides with the results in Butler and Loomes (2007). Even in the third round (see Table V), where the lower number of preference reversals could be interpreted as evidence of better formed preferences, the ratios were 18:0 (Bisection), 17:1 (Ping-pong) and 19:0 (List) in favour of the P-bet. That is, an impressive 54:1 in favour of the P-bet in valuation vs. choice. However, HCMB and RCB ratios were only 9:5 and 6:1, respectively (p=0.42 and p=0.13, McNemar exact binomial test, 2-sided) for a total of 15:6.

INSERT TABLE III

INSERT TABLE IV

INSERT TABLE V

In summary, for Bisection, Ping-pong and List, the results are in agreement with the evidence in Butler and Loomes (2007), that is, the P-bet is more highly preferred in matching than in choice.

The common element of the two methods where preference reversals almost disappeared (HCBM and RBC) is that they are non-transparent: one is iterative and the other is not. It seems that, when subjects are not aware that each choice is part of a sequence, they make each decision using the same principles that they use in straight choices. The method that leads to most discrepancies between valuation and choice is List. It seems that this presentation makes the matching attribute even more salient in the valuation task, and probabilities are even more overweighted. These results seem to support the Task Goal Hypothesis of Fischer et al. (1999) as an explanation of preference reversals between choice and matching.

We can see that the results are quantitatively (but not qualitatively) different for pairs A-B and C-D. There are more preference reversals in the pair C-D than in A-B. However, the results are qualitatively similar. Transparent methods produce a ratio 23:0 for pair A-B and 31:0 for pair C-D. Non-transparent methods produce a ratio 5:3 for pair A-B and 10:3 for pair C-D. Pair C-D clearly reproduces the pattern observed in Butler and Loomes (2007), namely, the P-bet more highly preferred in matching than in choice. This gives rise to a large number of preference reversals and it is in this context where the effect of non-transparent methods can be better observed. For pair A-B, we can also observe the pattern described in Butler and Loomes (2007) but to a lesser degree. For this reason, the "correcting" effect of non-transparent methods is less important for pair A-B.

4.2 Aggregate analysis

The preference reversal asymmetry is a phenomenon that applies to individuals. Therefore, we allocate individuals to one of eight groups according to the relationship between valuation and choice following the principles presented in the methods section (see 3.5.2.). Table VI shows the results for each valuation method and Table VII presents these results according to the degree of consistency of the subjects in their choices (strong vs. weak) and adding together the figures of transparent (Bisection, Ping-pong, List) and non-transparent (HCBM, RBC) methods. The main results are:

INSERT TABLE VI

INSERT TABLE VII

1. The relative number of preference reversals is different depending on the degree of consistency of subjects, which we take as a proxy for imprecision. Among subjects classified as "strongly consistent", in the case of pair A-B there are 17 participants out of 194 (8.8%), with 'true' preference reversal preferences. They prefer one lottery in each of the three choices but give a higher value to the other lottery in valuation. The ratio is 16:1 in the direction expected, namely, the P-bet more favoured in valuation than in choice. If we focus on those with less precise preferences (i.e. "weakly consistent"), in the case of pair A-B there are 15 participants out of 56 (26.8%), with 'true' preference reversal preferences (three times more than those "strongly consistent") and the asymmetry persists (12:3). In the case of pair C-D those percentages are 17.8% (ratio 34:0) for those "strongly consistent" and 37.3% for the "weakly consistent" (ratio 21:1). In summary, imprecision seems to have a very strong effect on preference reversals.

2. The List method stands out as the method that clearly produces most preference reversals. For all the 50 subjects who followed this procedure, the percentage of preference reversals is 20 % in pair A-B (ratio 10:0) and 38% in pair C-D (ratio 19:0). Even for the consistent subjects the percentages are very high, namely, 25% (ratio 9:0) and 34% (ratio 13:0), respectively for pairs A-B and C-D.
3. When we compare transparent (Bisection, Ping-pong and List) and non-transparent (HCBM and RBC) methods the frequency of preference reversals is larger for non-transparent methods. For the pair A-B, we have that 14.0% of subjects reverse their preferences in transparent methods and 11.0% in the case of non-transparent methods. For the pair C-D, those percentages are 29.3% for transparent methods and 12.0% for non-transparent. It is true that the transparent methods are penalized for the inclusion of the List method. If we only compare Bisection + Ping-Pong vs. HCBM + RBC, for the pair A-B the total number of subjects with preference reversal preferences is the same (i.e. 11) with the asymmetric ratio of 9:2 in both cases. No difference here. However, for the pair C-D, there is a clear difference. The total number of subjects with preference reversal preferences is 25 (ratio 25:0) for the transparent methods and 12 for non-transparent (ratio 11:1). This confirms that using non-transparent methods the number of preference reversals is reduced.

5. CONCLUSIONS

The main objective of this paper was to understand better the mechanisms that generate preference reversals and how to avoid/reduce that phenomenon. Our main result is that two explanations of preference reversals, presented at the beginning of

the paper, are complementary. First, subjects that are more consistent in repeated choices produce less preference reversals. As far as we know, this is the first paper to explore the role of imprecision as an explanation of preference reversals in health. However, even when we focus only on strongly consistent subjects, the asymmetric pattern of preference reversals continues when transparent methods are used. Methods based on non-transparent choices reduce the rate of preference reversals for both strongly and weakly consistent subjects. This seems to support the explanation of preference reversals based on the Task-Goal hypothesis. It is evident in our results that this finding is more compelling with pair C-D than with pair A-B, which suggests that the propensity for asymmetric reversals to occur may depend on the particular pairs involved. A key difference between pairs A-B and C-D is that, in straight choices, a majority of subjects (about 60%) chose the P-bet over the $-bet in the first case, whereas with pair C-D more than 60% of the sample chose the $-bet over the P-bet, which makes it easier to find the expected asymmetry. This shows that it is more difficult to investigate preference reversals when outcomes are health states than when they are money. If outcomes are monetary, it is easier to design lotteries with some common features, for example, with similar expected value. We have tried to devise lotteries with similar expected utilities, but we have used population averages in order to do that. However, the preferences of our subjects can be different from those averages. The important point is that in those pairs (like C-D) where preference reversals are more abundant, we have shown that both imprecision and the transparency of the method are relevant to explain preference reversals.

What are the implications of these results for preference elicitation methods in health? One is that matching methods should try to hide, as much as possible, the goal of the

task. In that way, the subject seems to treat each choice in the iteration process independently from the rest, without being influenced by past choices or other considerations. Furthermore, the results of HCBM suggest that it may not be necessary to abandon iterative methods and move to non-iterative ones, like RBC (that is, Discrete Choice Experiments), to avoid compatibility effects. This is an interesting finding since iterative methods are more efficient (require fewer questions) than non-iterative. Our results highlight the importance of the distinction between transparent and non-transparent methods. This distinction is not new, but we do not think that the vast majority of researchers who use matching methods, like Time Trade-Off or Standard Gamble, are aware of the potential relevance of using non-transparent methods. It is not uncommon to find papers arguing that they use CBM because "it leads to fewer inconsistencies than directly asking subjects for their indifference values" (Bleichrodt, Gao, & Rohde, 2016, p. 220) without mentioning the difference between transparent and non-transparent CBM. It is as if they assume that all CBM produce fewer inconsistencies. This paper suggests that this is not the case. Our results show that moving from standard matching to sequences of binary choices is not enough to avoid the problems of standard matching. If we want to use CBM methods to estimate utilities for health states, non-transparent methods seem to reduce biases present in transparent methods, reducing/avoiding compatibility effects and making choice and matching more similar.

The implication of our results is more complicated in the case of imprecision. While biases due to compatibility can be reduced using different techniques (i.e. non-transparent methods), imprecision cannot be addressed with such a "simple" change. The only solution here seems to be to help subjects to increase the degree of precision.

This is not easy in the case of the evaluation of health states. In spite of that, since preference imprecision seems to have an effect on how people respond to survey questions in the health domain, it is important to have an idea of the degree of imprecision involved in subject's responses.

ACKNOWLEDGEMENTS

The authors would like to give special thanks to Professor Graham Loomes for his suggestions about how to improve the paper, particularly in relation to the issue of imprecision. The first version of this paper did not pay much attention to this question and it was Professor Loomes who advised to reanalyse the data in order to include the role of preference imprecision. This turned out to be decisive in the explanation of the results of this paper. The authors also want to thank the reviewers for their suggestions, and express their gratitude to the Spanish Ministry of Economy, Industry and Competitiveness for its financial support (Projects ECO2013-43526-R / ECO2013-48631-P).

REFERENCES

Attema, A., & Brouwer, W. B. (2013). In search of preferred elicitation method: A test of the internal consistency of choice and matching tasks. *Journal of Economic Psychology, 39*,126-140.

Badia, X., Roset, R., Herdman, M., & Kind, P. (2001). A comparison of GB and Spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making, 21*(1), 7-16.

Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. (2012). Using a discrete choice experiment to estimate health state utility values. *Journal of Health Economics, 31*(1), 306-318.

Bleichrodt, H., Gao, Y., & Rohde, K. I. M. (2016). A measurement of decreasing impatience for health and money. *Journal of Risk and Uncertainty, 52*(3), 213-231.

Bostic, R., Herrstein, R.J., & Luce, R. (1990). The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior and Organization, 13*(2), 193-212.

Butler, D. J., & Loomes, G. C. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review, 97*(1), 277-297.

Fischer, G. W., & Hawkins, S. A. (1993). Strategy compatibility, scale compatibility and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance, 19*(3), 580-597.

Fischer, G. W., Carmon, Z., Ariely, D., & Zauberman, G. (1999). Goal-based construction of preferences: Task goals and the prominence effect. *Management Science, 45*(8), 1057-1075.

Fitts, P. M., & Seeger, C. M. (1953). S-R compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology, 46*(3), 199-210.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review, 92*(5), 1644-1655.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*(1), 46-55.

Loomes, G., & Pogrebna, G. (2016). Do preference reversals disappear when we allow for probabilistic choice? *Management Science, 63*(1), 163-184.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal, 92*(368), 805-824.

MacCrimmon, K., & Smith, M. (1986). *Imprecise equivalences: Preference reversals in money and probability*. Working paper, University of British Columbia, Vancouver, BC, Canada.

Oliver, A. (2013a). Testing procedural invariance in the context of health. *Health Economics, 22*(3)*,* 272-288.

Oliver, A. (2013b). Testing the rate of preference reversal in personal and social decision-making. *Journal of Health Economics, 32*(6)*,* 1250-1257.

Robinson, A., Spencer, A. E., Pinto-Prades, J. L., & Covey, J. A. (2016). Exploring Differences between TTO and DCE in the Valuation of Health States. *Medical Decision Making, 37*(3)*,* 273-284.

Seidl, C. **(**2002). Preference Reversal. *Journal of Economic Surveys*, 16(5): 621–55.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In R. M. Hogarth (Ed.), *Insights in Decision Making: A Tribute to Hillel J. Einhorn* (pp. 5-27). Chicago: The University of Chicago Press.

Sumner, W., & Nease, F. (2001). Choice-matching preference reversals in health outcome assessments. *Medical Decision Making, 21*(3), 208-218.

Tversky, . Sattath,S. & Slovic, P. (1988) Contingent weighting in judgment and choice. *Psychological Review*, 95 (3), 371-384.

TABLES

Table I. Lotteries used in the study.

	Pair 1	EU*	Pair 2	EU*
P-bet	**A**: (12231, 0.95; *Death*)	0.21	**C**: (22223, 0.8; *Death*)	0.12
$-bet	**B**: (11221, 0.3; *Death*)	0.24	**D**: (12221, 0.2; *Death*)	0.14

* Expected Utility according to Spanish tariff: U(12231)=0.219; U(11221)=0.816; U(22223)=0.141; U(12221)=0.682.

Table II. Reference gambles in Probability Equivalent questions.

	A: (12231, 0.95; D)	**B: (11221, 0.3; D)**	**C: (22223, 0.8; D)**	**D: (12221, 0.2; D)**
R:	(11111, 0.1; D)	(11111, 0.03; D)	(11111, 0.08; D)	(11111, 0.02; D)
	(11111, 0.2; D)	(11111, 0.06; D)	(11111, 0.16; D)	(11111, 0.04; D)
	(11111, 0.3; D)	(11111, 0.09; D)	(11111, 0.24; D)	(11111, 0.06; D)
	(11111, 0.4; D)	(11111, 0.12; D)	(11111, 0.32; D)	(11111, 0.08; D)
	(11111, 0.5; D)	(11111, 0.15; D)	(11111, 0.40; D)	(11111, 0.10; D)
	(11111, 0.6; D)	(11111, 0.18; D)	(11111, 0.48; D)	(11111, 0.12; D)
	(11111, 0.7; D)	(11111, 0.21; D)	(11111, 0.56; D)	(11111, 0.14; D)
	(11111, 0.8; D)	(11111, 0.24; D)	(11111, 0.64; D)	(11111, 0.16; D)
	(11111, 0.9; D)	(11111, 0.27; D)	(11111, 0.72; D)	(11111, 0.18; D)

Table III. Direct choices vs. choices implied by the valuation task. Round 1.

	Pair (A,B) n=225 [1]				Pair (C,D) n=225 [2]			
		Valuation		McNemar (p-value) [3]		Valuation		McNemar (p-value) [3]
	Choice	A≻B	B≻A		Choice	C≻D	D≻C	
Bisection	A≻B	24	4	0.7518	C≻D	15	1	0.0019
	B≻A	6	16		D≻C	14	20	
Ping-pong	A≻B	27	3	0.2278	C≻D	18	1	0.0019
	B≻A	8	12		D≻C	14	17	
HCBM	A≻B	25	4	0.7518	C≻D	11	2	0.0704
	B≻A	6	15		D≻C	9	28	
List	A≻B	17	0	0.0002	C≻D	10	0	<0.0001
	B≻A	16	11		D≻C	25	9	
RBC	A≻B	12	0	0.0412	C≻D	7	2	0.6831
	B≻A	8	11		D≻C	6	16	

[1] Due to inconsistencies in valuation, it was not possible to obtain the indifference interval in 6 cases in the List group and in 19 subjects in the RBC group. [2] The indifference interval could not be identified in 6 occasions in the List group and in 19 cases in the RBC group. [3] McNemar's test 2-sided.

Table IV. Direct choices vs. choices implied by the valuation task. Round 2.

	Pair (A,B) n=235 [1]				Pair (C,D) n=239 [2]			
		Valuation		McNemar (p-value) [3]		Valuation		McNemar (p-value) [3]
	Choice	A≻B	B≻A		Choice	C≻D	D≻C	
Bisection	A≻B	28	2	0.2888	C≻D	19	0	0.0009
	B≻A	6	14		D≻C	13	18	
Ping-pong	A≻B	35	1	1.0000	C≻D	22	0	0.0026
	B≻A	0	14		D≻C	11	17	
HCBM	A≻B	28	2	1.0000	C≻D	14	2	0.6831
	B≻A	3	17		D≻C	4	30	
List	A≻B	26	0	0.0233	C≻D	18	0	0.0003
	B≻A	7	8		D≻C	15	13	
RBC	A≻B	23	0	0.2420	C≻D	15	1	0.3711
	B≻A	3	18		D≻C	4	23	

[1] Due to inconsistencies in valuation, it was not possible to obtain the indifference interval in 9 cases in the List group and in 6 subjects in the RBC group. [2] The indifference interval could not be identified in 4 occasions in the List group and in 7 cases in the RBC group. [3] McNemar's test 2-sided.

Table V. Direct choices vs. choices implied by the valuation task. Round 3.

	Pair (A,B) n=236 [1]				Pair (C,D) n=228 [2]			
		Valuation		McNemar (p-value) [3]		Valuation		McNemar (p-value) [3]
	Choice	A≻B	B≻A		Choice	C≻D	D≻C	
Bisection	A≻B	30	0	0.0133	C≻D	20	0	0.0044
	B≻A	8	12		D≻C	10	20	
Ping-pong	A≻B	31	1	0.0771	C≻D	23	0	0.0044
	B≻A	7	11		D≻C	10	17	
HCBM	A≻B	29	2	0.6171	C≻D	12	3	0.3428
	B≻A	2	17		D≻C	7	28	
List	A≻B	25	0	0.0133	C≻D	15	0	0.0026
	B≻A	8	9		D≻C	11	12	
RBC	A≻B	25	1	0.6171	C≻D	17	0	0.2482
	B≻A	3	15		D≻C	3	20	

[1] Due to inconsistencies in valuation, it was not possible to obtain the indifference interval in 8 cases in the List group and in 6 subjects in the RBC group. [2] The indifference interval could not be identified in 12 occasions in the List group and in 10 cases in the RBC group. [3] McNemar's test 2-sided.

Table VI. Analysis of Preference Reversals at the individual level (n=250).

	Choice	Valuation A≻B	Valuation B≻A	McNemar (p-value) [1]	Choice	Valuation C≻D	Valuation D≻C	McNemar (p-value) [2]
Bisection	3A	23	1	0.4497	3C	12	0	0.0015
	2A	5	1		2C	7	0	
	2B	3	1		2D	5	0	
	3B	2	14		3D	7	19	
Ping-pong	3A	27	0	0.1336	3C	15	0	0.0009
	2A	6	0		2C	6	0	
	2B	3	2		2D	6	1	
	3B	1	11		3D	7	15	
HCBM	3A	26	0	0.3711	3C	9	0	0.1306
	2A	1	1		2C	4	1	
	2B	3	5		2D	3	4	
	3B	1	13		3D	3	26	
List	3A	18	0	0.0044	3C	13	0	<0.0001
	2A	13	0		2C	6	0	
	2B	1	0		2D	6	0	
	3B	9	9		3D	13	12	
RBC	3A	20	0	0.2207	3C	16	0	0.0736
	2A	7	1		2C	4	0	
	2B	2	1		2D	1	5	
	3B	3	16		3D	4	20	

Shaded cells show individuals who are "strongly consistent", in the sense that they always prefer the same lottery in all the three straight choices.

[1] McNemar's test 2-sided. Figures in rows 3A and 2A have been merged, as well as rows 2B and 3B, to obtain the p-values. [2] McNemar's test 2-sided. Figures in rows 3C and 2C have been merged, as well as rows 2D and 3D, to obtain the p-values.

Table VII. Preference Reversals in strongly and weakly consistent subjects [*].

		Choice	Valuation A≻B	Valuation B≻A	Rate of PR (%)		Choice	Valuation C≻D	Valuation D≻C	Rate of PR (%)
Transparent [1]	Strong	A≻B	68	1	11.3	Strong	C≻D	40	0	23.9
		B≻A	12	34			D≻C	27	46	
	Weak	A≻B	24	1	22.9	Weak	C≻D	19	0	45.9
		B≻A	7	3			D≻C	17	1	
	Total	A≻B	92	2	14.0	Total	C≻D	59	0	29.3
		B≻A	19	37			D≻C	44	47	
Non-transparent [2]	Strong	A≻B	46	0	5.1	Strong	C≻D	25	0	9.0
		B≻A	4	29			D≻C	7	46	
	Weak	A≻B	8	2	33.3	Weak	C≻D	8	1	22.7
		B≻A	5	6			D≻C	4	9	
	Total	A≻B	54	2	11.0	Total	C≻D	33	1	12.0
		B≻A	9	35			D≻C	11	55	
TOTAL	Strong	A≻B	114	1	8.8	Strong	C≻D	65	0	17.8
		B≻A	16	63			D≻C	34	92	
	Weak	A≻B	32	3	26.8	Weak	C≻D	27	1	37.3
		B≻A	12	9			D≻C	21	10	

[*]See main text for definitions.

[1] Transparent methods include Bisection, Ping-pong and List. [2] Non-transparent methods include HCBM and RBC.

FIGURES

Figure 1. Example of a direct choice task (Lottery A vs. Lottery B).

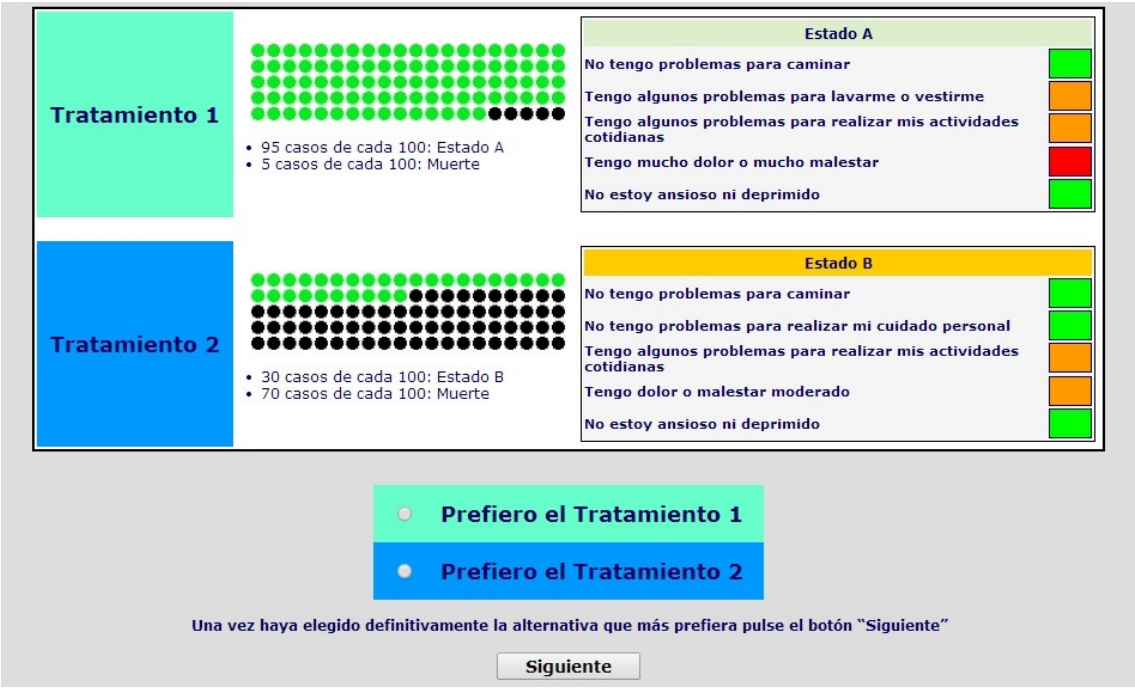


Figure 2. Example of a RPE question.

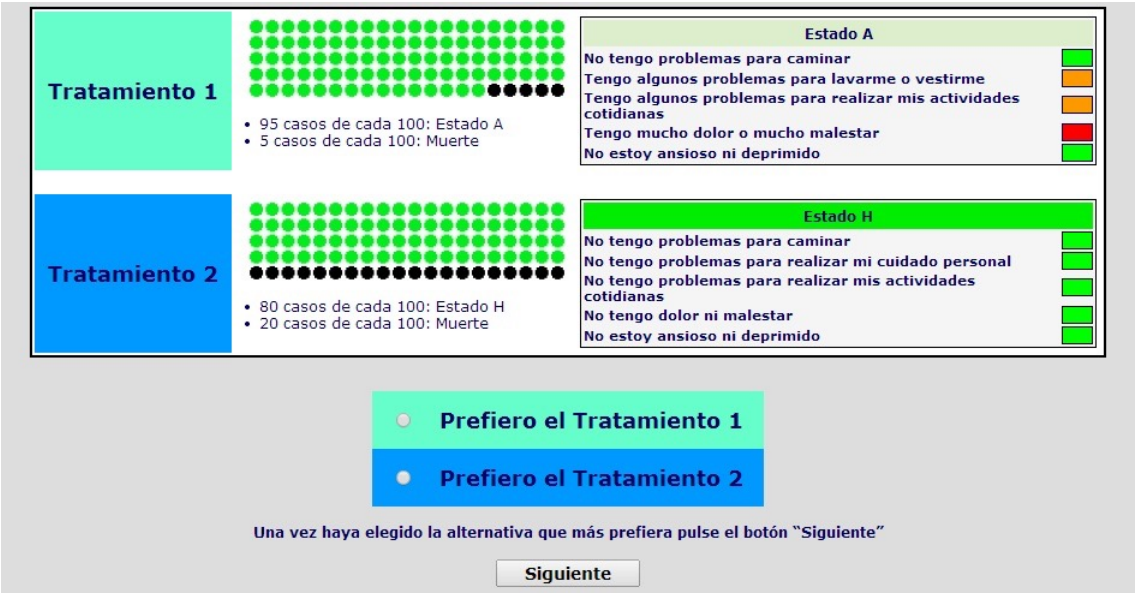
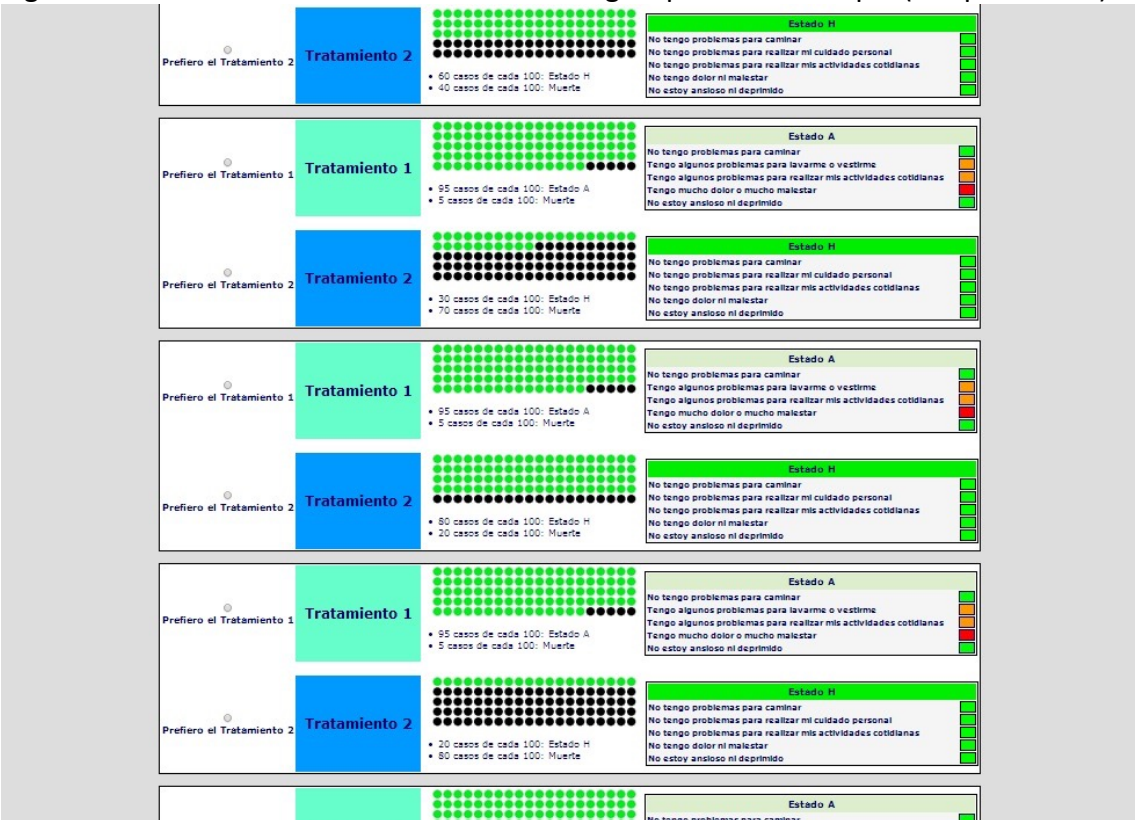


Figure 3. Detail of a set of choices in a matching sequence in Group 4 (*List* procedure).



APPENDIX

Results assuming that a preference is implied by the valuation task (RPE) when intervals do not overlap

Table A1. Direct choices vs. choices implied by the valuation task. Round 1.

	Choice	Valuation A≻B	Valuation B≻A	Valuation A≈B[1]	McNemar (p-value) [2]	Choice	Valuation C≻D	Valuation D≻C	Valuation C≈D[1]	McNemar (p-value) [2]
Bisection	A≻B	23	0	5	0.1336	C≻D	15	1	0	0.0159
	B≻A	4	8	10		D≻C	10	15	9	
Ping-pong	A≻B	26	2	2	0.2888	C≻D	18	0	1	0.0026
	B≻A	6	8	6		D≻C	11	15	5	
HCBM	A≻B	24	0	5	0.2482	C≻D	10	2	1	0.4497
	B≻A	3	11	7		D≻C	5	22	10	
List	A≻B	16	0	1	0.0015	C≻D	10	0	0	<0.0001
	B≻A	12	9	6		D≻C	19	7	8	
RBC	A≻B	11	0	1	0.2482	C≻D	7	2	0	1.0000
	B≻A	3	11	3		D≻C	0	14	7	

[1] Cases in which preference could not be inferred from the RPE responses, since indifference intervals for each lottery overlap. [2] McNemar's test 2-sided.

Table A2. Direct choices vs. choices implied by the valuation task. Round 2.

	Choice	Valuation A≻B	Valuation B≻A	Valuation A≈B[1]	McNemar (p-value) [2]	Choice	Valuation C≻D	Valuation D≻C	Valuation C≈D[1]	McNemar (p-value) [2]
Bisection	A≻B	27	0	3	0.0736	C≻D	18	0	1	0.0736
	B≻A	5	13	2		D≻C	5	13	13	
Ping-pong	A≻B	34	1	1	1.0000	C≻D	22	0	0	0.0077
	B≻A	0	8	6		D≻C	9	16	3	
HCBM	A≻B	27	2	1	1.0000	C≻D	12	1	3	0.4795
	B≻A	3	13	4		D≻C	1	23	10	
List	A≻B	26	0	0	0.0736	C≻D	17	0	0	0.0026
	B≻A	5	7	3		D≻C	11	9	8	
RBC	A≻B	23	0	0	1.0000	C≻D	14	1	0	1.0000
	B≻A	1	16	4		D≻C	0	17	10	

[1] Cases in which preference could not be inferred from the RPE responses, since indifference intervals for each lottery overlap. [2] McNemar's test 2-sided.

Table A3. Direct choices vs. choices implied by the valuation task. Round 3.

	Choice	Valuation			McNemar (p-value) [2]	Choice	Valuation			McNemar (p-value) [2]
		A≻B	B≻A	A≈B[1]			C≻D	D≻C	C≈D[1]	
Bisection	A≻B	29	0	1	0.4795	C≻D	20	0	0	0.0233
	B≻A	2	9	9		D≻C	7	12	11	
Ping-pong	A≻B	29	0	3	0.0736	C≻D	23	0	0	0.0233
	B≻A	5	7	6		D≻C	7	15	5	
HCBM	A≻B	27	2	2	0.4795	C≻D	12	2	1	0.6831
	B≻A	0	17	2		D≻C	4	24	7	
List	A≻B	25	0	0	0.0233	C≻D	14	0	1	0.0077
	B≻A	7	7	3		D≻C	9	9	5	
RBC	A≻B	25	1	0	1.0000	C≻D	17	0	0	1.0000
	B≻A	0	12	5		D≻C	0	14	9	

[1] Cases in which preference could not be inferred from the RPE responses, since indifference intervals for each lottery overlap. [2] McNemar's test 2-sided.